Repeated Bisections Approach for Local Clustering of PPINs

Zakharia Frenkel[1], Sandeep Amberkar[2], Lars Kaderali[3], Zeev Volkovich[4]

[1]Genome Diversity Center, Institute for Evolution, University of Haifa, Haifa, Israel

[1,4]Department of Software Engineering, ORT Braude College, Karmiel, Israel

[2]Faculty of Biosciences, University of Heidelberg, BW, Germany

[2,3]Institute for Medical Informatics and Biometry (IMB), University of Technology Dresden, Medical School, Dresden, Germany

[1]zakharf@research.haifa.ac.il; [2]sandeep.amberkar@bioquant.uni-heidelberg.de; [3]lars.kaderali@bioquant.uni-heidelberg.de; [4]vlvolkov@braude.ac.il

Abstract

In this paper we introduce a new heuristic approach for local clustering of the protein-protein interaction networks (PPIN), which can be applied to very large graphs. The method is based on idea of repeated bisections (rbr) proposed earlier for global clustering of PPIN. Each round of bisection is carried out by multilevel graph clusterization method realized by "Graculus" tool.

Keywords

Protein-protein Interaction Network; Local Clustering; Repeated Bisections; "Graculus" Tool

Introduction

The network approach is very popular in many branches of the modern science and engineering. Usually, it is applied for description of systems composed of a large number of connected components. The term "connection" here can reflect physical association, ability to transmit something (for instance, information), similarity, and so on. Such prevalence and requirement for the networks has stimulated development of a large variety of computational algorithms and analytical approaches for large network access, navigation and analysis. Graph based clustering is one of the typical tasks of the network analysis. This task can be defined as assigning a set of objects (i.e. vertices, in our case) into groups (i.e. clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters.

There are a lot of algorithms for cluster analysis based on different measures of similarity and different strategies for achievement of the optimal grouping. These measures of similarity called also graph clustering objectives. Commonly used objectives are normalized cut, ratio association value, and others. Such a wide variety of approaches results from the fact that network with different structures (i.e. built for different objectives) requires different measures and strategies for optimal solution.

In this paper we describe a new approach which was applied to clustering of the protein-protein interaction networks (PPIN). The PPINs reflect application of integrative approach to understanding of functionality of very complicated biological systems composed of a large number of components (largely proteins, but may also include genes, RNAs and other molecules) that can interact with each other by rather complicated ways. PPINs were successfully applied to prediction of a single protein function, signaling and metabolic pathways, as well as to understanding of physiological processes and molecular basis of some diseases. The clustering of the PPINs is of particular importance as it makes possible to identify protein complexes and functional modules. The latter aspect of PPINs is very essential for functional annotation of proteins. According to its importance the PPIN clustering is presently a well-developed field, however, the problem is far from a satisfactory solution.

Generally, the multitude of clustering algorithms (unfortunately, in the literature they, indeed, called by the same term 'clustering') can be divided into two groups: global clustering (i.e. assigning of each vertex into a group) and local clustering, where only a set of subgraphs of required properties is selected. Most of the algorithms developed for PPIN are directed to a solution of local clustering. One of common clustering criteria in this case is density, which can be defined by different ways. The simple 'density' approach (where the density is defined by subgraph closeness of the clique) has significant limitations. Many alternative approaches were proposed since that time. Generally, a better quality of clustering equals to more algorithm complexity, which makes it difficult to find appropriate solution for huge PPINs. It is well known that the clique finding problem is NP-hard. If the cliques are not 'complete' the complexity of the problem increases by many-fold.

In this paper we describe a new, very simple yet effective heuristic algorithm for local clustering. Actually, our approach is a modification of repeated bisections (rbr) algorithm used for global clustering of PPIN. For bisection at every step we used "Graculus" tool of Dhillon and co-workers, which is very quick and able to work on large graphs. The main advantage of their multilevel approach is that it decreases the graph cut objective even at the refinement stage. This makes the algorithm much more effective than previous versions developed in the field.

Methods

Algorithm Description

As it was mentioned above, our algorithm is related to the repeated bisections' methodology (rbr). In its original form, the rbr is a top-down clustering algorithm, with input of desired number of clusters (k). Correspondingly, k−1 repeated bisections are carried out, and k output clusters are obtained. In our case, when the purpose is local clustering of data, these bisections are continued until some user-defined criterion has been met. The algorithm works recursively (function F (G)), iterating at every level:

1) Analyse the input graph G and check its correspondence to a specified condition (we selected density definition as (2*m/ (n*(n-1)), where m – number of edges, n – number of vertexes), but another definition also can be used).

If the graph satisfies the condition or its size less than the selected value, the function returns the input graph G; Else:

2) Divide the input graph G into two clusters (sub-graphs) G1 and G2 by Dihlon algorithm (see below) with selected objective function.

3) For obtained sub-graphs G1 and G2 run F (G1) and F (G2).

Generally, for some cases, the "Graculus" tool which we used for bisection can produce more than two connected components after the graph partition on two clusters. For each connected component F(G) function should be run. To decrease amount of connected components in the new graphs, we deleted all "side branches" of the networks at each step (i.e. the nodes that are not belong to cycles or connecting paths between cycles; see Fig. 1).

"Graculus" Tool

For graph bisection at each stage we use Graclus software. Description and code is available at http://www.cs.utexas.edu/users/dml/Software/graclus.html. The algorithm proposed by of Dhillon and coworkers is multilevel, it includes three stages: coarsening phase, initial clustering phase (which carried out by spectral algorithm) and refinement phase. The authors have shown that (with correct choice of weights and kernel matrix) the weighted kernel k-means algorithm can be directly used to locally optimize all commonly used graph clustering objectives. So, at each stage of refinement "Graculus" applies (improved by a local search algorithm considers the effect on the objective function of moving a point from one cluster to another) the weighted kernel k-means algorithm with initial partition obtained from clustering of the previous level.

We used two objectives for optimization during the graph clustering: normalized cut (ncut) and ratio association (rassoc). The first defined as the cut relative to the degree of a cluster NCUT = links (Vc,V\Vc) / degree (Vc). The second is within-cluster association relative to the size of the cluster: RASSOC = links (Vc,Vc) / |Vc|. The task of graph clustering is to find minimum of NCUT and maximum of RASSOC.

Test Networks

Protein interactions were collated from a meta-database called iRefindex which itself has interactions from several public repositories such as DIP, MINT, IntAct, CORUM, HPRD, MPPI, OPHID, BioGRID and BIND. In addition, high confidence (confidence score greater than 0.750) interactions from STRING were included. This dataset was called HSA.Unified.Interactome.750 or HSA750. We then extracted a subset of interactions from this dataset with confidence score greater than 0.900. This dataset was called HSA.UnifiedInteractome.900 or HSA900. The HSA750 has 18941 proteins with 377961 interactions between them while HSA900 has 17764 proteins with 181870 interactions between them. These interactions didn’t include self-loops, but the networks contained more than one connected components.

Results and Discussion

For application of the clustering algorithms HSA_Unified Interactome networks 900 (see "Methods") was selected. The main isolated component contains 17208 vertices (118453 edges) correspondingly. The most of such brunches are single points, so this procedure significantly decreases amount of vertices (13808) without significant change of edges number (115053). The graph is relatively dense; and vertices strength distribution is shown in Fig. 2.

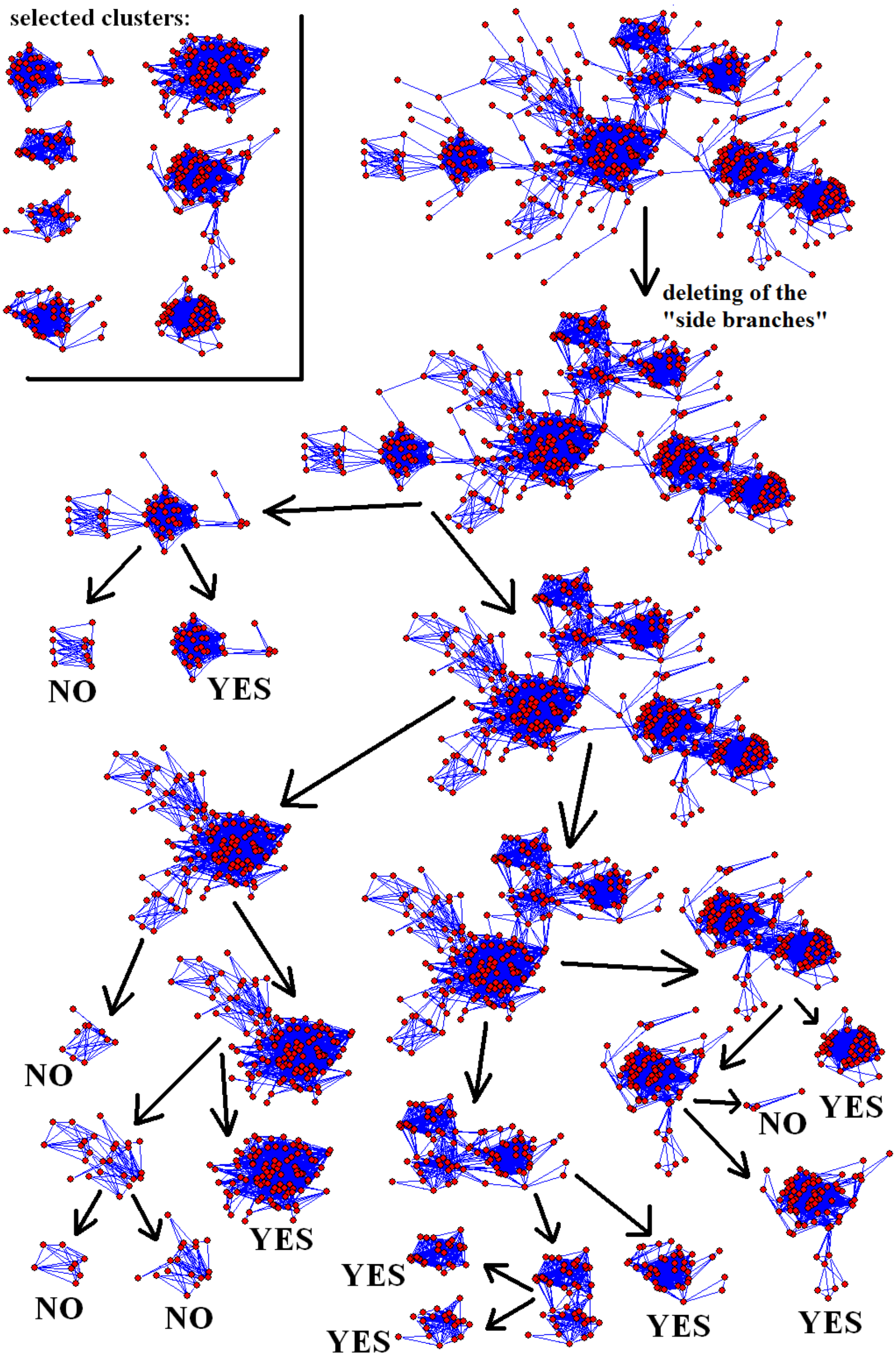

FIG. 1 ALGORITHM SCHEME ARROWS SHOW TO RECURSIVE BISECTION. "YES" – MEANS THAT THE OBTAINED CLUSTER IS SATISFIED TO REQUIRED CRITERION; "NO"- MEANS THAT THE GRAPH IS TOO SMALL, AND LOCAL SEARCH IN THIS DIRECTION IS FINISHED

We apply our bisection algorithm for dense groups searching in this huge graph for different density threshold. A simple definition of cluster density was used: d = 2*m/ (n*(n-1), i.e. closeness of the cluster to the clique. The resulting dependences of obtained amount of clusters and average cluster size on density thresholds are shown in Figs. 3A and 3B, correspondingly. Algorithm was used with number minimal cluster size – five nodes. Two graph cut objectives were used: normalized cut (ncut) and ratio association (rassoc), but results were not significantly different.

The curve of clusters amount dependence on density thresholds has a bell-shaped form (Fig. 3A). It means that at lower density requirements, many nodes are united to form several large sparse clusters. Subsequent increase in the threshold leads to decomposition of the sparse large clusters into the multiple smaller denser clusters. At the highest thresholds, these smaller clusters are also decomposed into groups smaller than selected threshold of 5 nodes for the cluster size. This explanation is consistent with monotonic decreasing of the average cluster size with increasing of the density threshold (Figs. 3B). Presumably, a curve maximum at position 0.25 (Fig. 3A) can point to optimal cluster density for selected type of graphs.

In general, the method can be also used for global clusterization. In such implementation, it requires to store the small clusters and to attach back deleted "side branches". The advantage of such global clusterization that it doesn't need for input of desired amount of clusters, whose selection often is not trivial task. The output clusters will be composed of the cores with density more than selected threshold (or size less than selected threshold) and "side branches".

The main advantage of the method is high speed of calculation that allows working with large graphs and carrying out clustering procedure repeatedly under different conditions. The second advantage is that it easily allows changing objective function (i.e. criterion) of clustering. However our approach also has two weaknesses. The first is that the exact boundaries of the clusters are often not optimal (see Fig. 1). This drawback can be easily corrected by using one of many developed refinement procedures for boundaries of each cluster relatively to the primary graph. The second disadvantage is with some probability we can "destroy" a good cluster during the repeated bisections, which would lead to lucking of its detection. A possible approach to solve this problem is introduction of some stochastic parameter into bisection and to carry out the procedure several times.

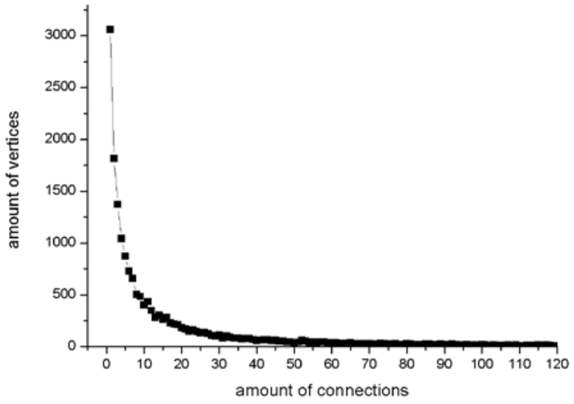

FIG. 2 NETWORK HSA 900

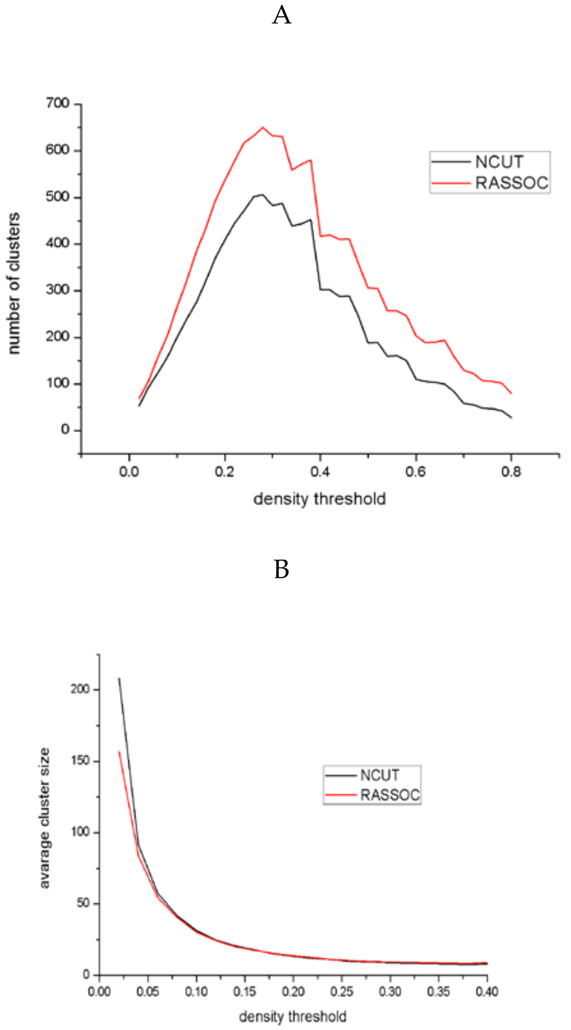

FIG. 3 CLUSTERING BY DICHOTOMY FOR DIFFERENT DENSITY THRESHOLDS FOR CLUSTER SELECTION A. NUMBER OF CLUSTERS; B. AVERAGE CLUSTER SIZE. THE MINIMAL PERMITTED CLUSTER SIZE IS FIVE NODES

ACKNOWLEDGMENT

This work has been supported by the European Union seventh framework program via the PathoSys project (grant number 260429).

REFERENCES

Altaf-Ul-Amin, Md, Shinbo, Yoko, Mihara, Kenji et al., "Development and implementation of an algorithm for detection of protein complexes in large interaction networks," Bmc Bioinformatics vol. 7, Apr 14 2006. pp. 2006.

Aranda, B., Achuthan, P., Alam-Faruque, Y. et al., "The IntAct molecular interaction database in 2010," Nucleic Acids Research vol. 38, Jan. pp. D525-D531, 2010.

Bader, G. D., Donaldson, I., Wolting, C. et al., "BIND - The Biomolecular Interaction Network Database," Nucleic Acids Research vol. 29 (1), Jan 1 2001. pp. 242-245, 2001.

Brown, K. R. and Jurisica, I., "Online predicted human interaction database," Bioinformatics vol. 21 (9), May 1 2005. pp. 2076-2082, 2005.

Ceol, Arnaud, Aryamontri, Andrew Chatr, Licata, Luana et al., "MINT, the molecular interaction database: 2009 update," Nucleic Acids Research vol. 38, Jan. pp. D532-D539, 2010.

Chautard, E., Thierry-Mieg, N., and Ricard-Blum, S., "Interaction networks: From protein functions to drug discovery. A review," Pathologie Biologie vol. 57 (4), Jun 2009. pp. 324-333, 2009.

Dhillon, I., Guan Y., and Kulis, B., presented at the Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2005.

Dhillon, Inderjit S., Guan Yuqiang, and Kulis, Brian, "Weighted graph cuts without eigenvectors: A multilevel approach," Ieee Transactions on Pattern Analysis and Machine Intelligence vol. 29 (11), Nov 2007. pp. 1944-1957, 2007.

Dhillon, I.S., Guan Y., and Kogan, J., presented at the Proceedings - IEEE International Conference on Data Mining, ICDM, 2002.

Dhillon, I.S., Guan Y., and Kulis, B., presented at the KDD-2004 - Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004.

Hastad, J., "Clique is hard to approximate within n(1-epsilon)," Acta Mathematica vol. 182 (1), 1999. pp. 105-142, 1999.

http://en.wikipedia.org/wiki/Cluster_analysis, vol.

Kaufman, L. and Rousseauw, P.J., "Finding groups in data. An introduction to cluster analysis". Wiley, New York, 1990.

Keshava Prasad, T. S., Goel, Renu, Kandasamy, Kumaran et al., "Human Protein Reference Database-2009 update," Nucleic Acids Research vol. 37, Jan 2009. pp. D767-D772, 2009.

Li Min, Chen Jian-er, Wang Jian-xin et al., "Modifying the DPClus algorithm for identifying protein complexes based on new topological structures," Bmc Bioinformatics vol. 9, Sep 25, 2008.

Pagel, P., Kovac, S., Oesterheld, M. et al., "The MIPS mammalian protein-protein interaction database," Bioinformatics vol. 21 (6), Mar 15 2005. pp. 832-834, 2005.

Razick, Sabry, Magklaras, George, and Donaldson, Ian M., "iRefIndex: A consolidated protein interaction database with provenance," Bmc Bioinformatics vol. 9, Sep 30 2008. pp. 2008.

Ruepp, Andreas, Brauner, Barbara, Dunger-Kaltenbach, Irmtraud et al., "CORUM: the comprehensive resource of mammalian protein complexes," Nucleic Acids Research vol. 36, Jan 2008. pp. D646-D650, 2008.

Salwinski, L., Miller, C. S., Smith, A. J. et al., "The Database of Interacting Proteins: 2004 update," Nucleic Acids Research vol. 32, Jan 1 2004. pp. D449-D451, 2004.

Schaeffer, S.E., "Graph clustering," Computer Science Review vol. 1 (1), 2007. pp. 27-64, 2007.

Spirin, V. and Mirny, L. A., "Protein complexes and functional modules in molecular networks," Proceedings of the National Academy of Sciences of the United States of America vol. 100 (21), Oct 14 2003. pp. 12123-12128, 2003.

Stark, Chris, Breitkreutz, Bobby-Joe, Chatr-aryamontri, Andrew et al., "The BioGRID Interaction Database: 2011 update," Nucleic Acids Research vol. 39, Jan. pp. D698-D704, 2011.

Szklarczyk, Damian, Franceschini, Andrea, Kuhn, Michael et al., "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored," Nucleic Acids Research vol. 39, Jan. pp. D561-D568, 2011.

Wang Jianxin, Li Min, Deng Youping et al., "Recent advances in clustering methods for protein interaction networks," Bmc Genomics vol. 11, Dec. pp. 2010.

Zhang Y., Zeng E., Li T. et al., presented at the 8th International Conference on Machine Learning and Applications, ICMLA 2009.